Concept Paper with Empirical Evidence

The Borrowed Warrant

How a System's Own Encodings Acquire the Standing of Results

A second mechanism at the same site as the capability boundary, inside it rather than at it, and why it is cheaper to control than the first

Paper IV of the Capability Boundary Papers

Nico Andreas Heller, Berlin, 17 August 2026

Abstract

The Capability Boundary Principle (Heller, 2026a) established that an artificial system asked to perform an operation beyond its architecture does not fail visibly: it performs a substitute operation and describes the result in the vocabulary the request supplied. That mechanism -- the borrowed voice -- masks an operation. This paper identifies a second mechanism at the same site -- the collaboration between a system and the source it is asked to work from -- which masks something else: a reading. The second mechanism does not cross the capability boundary at all: every operation in it, and every check that would expose it, lies within the system's own register. It is a failure inside the boundary rather than at it, and Paper I §6 now says so. A system encodes its interpretation of a source as machine-readable data; an instrument executes the encoding; the execution's output is reported as a computation over the source; and the report is subsequently cited as evidence for the interpretation. Every step is individually valid. The loop closes, and an interpretation acquires the standing of a result. The paper calls this the borrowed warrant, establishes it by intervention rather than by observation -- each load-bearing claim is tested by removing its encoded support and observing whether the claim survives -- and reports five instances from a software-architecture production programme in which everyone passed a purpose-built verification suite that reported itself valid throughout. The paper distinguishes the mechanism from five near neighbours by stated tests, and identifies an asymmetry with practical consequences: detecting a borrowed voice requires a capability the system lacks, whereas detecting a borrowed warrant requires only a position the system cannot occupy. The second is therefore far cheaper to control, and is at present not controlled at all. A falsifiable prediction is given. The paper also applies its own standard to itself: three of its quantified consequences are withdrawn because the instruments that produced them are no longer retained, and a cheap screen it had offered as corroboration is shown, by the ablation the paper recommends, to rest on an undisclosed hand-authored surface and not to survive it. Both corrections were produced by the controls proposed here, applied to this paper.

Contact: n.a.heller@democracyschool.com

1. Two ways to be wrong while being correct

A system is given a formal specification and asked to derive an architecture from it. It reads one sentence of that specification and encodes its reading as a machine-readable relation: *this element brings about that one*. The relation joins a set of others. An instrument executes the set and emits an ordering. The ordering is reported as what the specification requires, and a design document is then built on it, citing the execution as its ground.

Consider what an auditor would find. The quotation attached to the encoded relation is present at the cited line, verbatim, and a mechanical check confirms it. The instrument's code is correct and deterministic. The test suite is green and says so explicitly. The design document's citations resolve. Nothing an ordinary review would examine is wrong.

Now look at what the relation asserts. Its target is an entire layer of the architecture. The cited sentence, read in full, does not mention that layer, and no sentence in the specification connects the two. Remove the relation and the instrument has no encoded warrant for the layer at all -- nor for the elements that hang beneath it. The layer was in the architecture because the system had put it there, and the system had put it there while reading.

(An earlier draft of this paper quantified that removal -- so many elements reached, so many lost. The instrument that produced those counts is no longer in the released set, and the figures are withdrawn under the paper's own retention rule, §10.3 and §3.6. The three facts above are readable in the artefacts as released, and the third was established by a party with no stake in the answer.)

The hinge is this. *The system did not read the specification and conclude. It read an artefact of its own making, in which a reading of the specification had been given the form of data, and concluded from that.*

This paper names that mechanism the borrowed warrant, and sets it beside the mechanism Paper I identified. The borrowed voice borrows *vocabulary*, and it borrows it *from the request*: asked for an operation it cannot perform, a system performs one it can and describes the result in the requester's terms, because those are the only terms available (Heller, 2026a, §1). The borrowed warrant borrows *warrant*, and it borrows it *from the system's own prior artefact*.

Both produce output that satisfies every surface criterion. They differ in one respect, and that difference governs everything downstream: what it takes to see them. Detecting a borrowed voice requires perceiving that the operation performed differs

categorically from the operation described -- which requires the very capability the substitute operation stood in for. Detecting a borrowed warrant requires tracing where an ordering came from, which is ordinary chain-tracing, well within the reach of the system that committed it. What the committing system cannot supply is not the capability. It is a position: an assessment of its own prior output, with no stake in the answer.

Positions are cheap.

2. The mechanism

2.1. The cycle

Let *S* be a source: a specification, a standard, a body of law, a dataset schema -- any text an artificial system is asked to work from rather than to write.

A reading **r** of *S* is encoded as datum **d**. An instrument **I** executes **d** and emits output **o**. Output **o** is reported as a computation over ***S***. Report ***o*** is then cited as evidence for reading ***r***.

The cycle can iterate. Each pass raises *r*'s standing: cited once as an encoding, *r* becomes, three artefacts later, a property of *S* that subsequent work is built upon.

2.2. No step is a fallacy

This must be conceded fully and early, because the mechanism is otherwise mistaken for carelessness, and remedies aimed at carelessness do not touch it.

The encoding is legitimate. Formalizing an interpretation is what makes it testable; a project that refused to encode readings could not compute at all.

The execution is legitimate. It is deterministic, inspectable, and does exactly what its code says. The report is accurate. It describes the execution correctly. It does not overstate what ran. The citation is accurate. The execution did occur, and did produce that output.

The defect is in the closure, and closure is a property of the cycle rather than of any link in it. No individual inspection finds it, because no individual link is defective.

2.3. Three properties that make it invisible

Surface-criterion completeness. Every check a competent reviewer would apply passes. Citation present. Quotation verbatim. Code correct. Tests green. Documentation exists. The set of things that would ordinarily raise suspicion is empty.

Provenance flattening. In the record, an instrument's output and a derivation from a source are the same *kind of object*: a reported result. The artefact carries no field distinguishing *this was computed from the source* from *this was computed from my reading of the source*. Once flattened, the distinction cannot be recovered by reading the artefact; it can only be recovered by re-running the instrument without the encoding.

Authority accrual. Standing rises monotonically with each citation. Nothing in the cycle reduces it, and the artefacts that cite the raised standing are themselves correct.

2.4. Why it is not detectable from inside

Not because the detecting operation lies beyond the system's register. Tracing an ordering to its source, and testing whether a conclusion survives the removal of its support, are sequential, decompositional operations -- the class of operations at which such systems are strongest (Heller, 2026a, §4).

The obstacle is structural rather than architectural. The party that would perform the check is the party that authored the encoding, and the check's outcome bears on its own prior output. This is not a claim about motivation, and §7 defends it against that reading. It is a claim about the relation between an assessor and the thing assessed: the encoding is not visible *as* an encoding to the process that produced it, because to that process it is simply what the source says.

3. Study design and the standard of evidence

This section is placed before the evidence because the evidence is worthless without it, and because the objections it answers are the ones a reader should raise.

3.1. What kind of study this is

A mechanistic critical case. The claim is that a mechanism exists, operates under stated conditions, and evades a stated class of controls. It is not a claim about how often the mechanism occurs, in what proportion of systems, or with what effect size.

The evidence appropriate to a mechanism is intervention, not frequency. A single well-instrumented case in which a mechanism is switched off and the dependent phenomenon disappears establishes the mechanism. A thousand observations in which it is never switched off establish only a correlation.

3.2. The inferential move

Every load-bearing claim in §5 is established by ablation:

Remove the encoded element. Re-run the unmodified instrument. Observe whether the conclusion survives.

A conclusion that disappears when its encoding is removed was carried by the encoding. A conclusion that survives was carried by something else. This is knockout logic, and its validity does not depend on the number of cases; it depends on the intervention being clean, the instrument being unmodified, and the run being deterministic. All three are met and are reproducible by a third party from the artefacts.

3.3. What the design can and cannot support

Can: that the mechanism exists; that it operated in this case; that it was invisible to every control applied to it, including controls purpose-built to prevent related failures; that it is detectable by a stated procedure at stated cost.

Cannot: prevalence in any population; generality across model architectures; effect size; whether the mechanism is more or less common than the borrowed voice.

The paper does not reclaim in later sections what it disclaims here.

3.4. The independence problem, and what was done about it

The analytical instrument that produced this evidence was built by the system whose failures the evidence reports, and the first interpretation of the evidence was that system's own. That is a disqualifying conflict for the paper's central question -- *did this ordering originate with the source or with the system?* -- because the question is precisely about the system's own contribution.

The response was a pre-registered adversarial review, conducted by a second instance of the same architecture:

- the question set was fixed before the reviewer saw any artefact;
- the authoring system's own account of what it had done was excluded by name, document by document, rather than by general instruction;
- the reviewer worked from primary artefacts and executable code, restored the instrument and ran it, and reproduced every run verbatim in appendices;
- the prompt stated explicitly that finding nothing was an acceptable and expected outcome, because a reviewer instructed to find problems finds them.

The reviewer returned findings against the instrument its own architecture had built, including one the authoring system would have defended. It also declined to perform a required session-opening procedure on discovering that the procedure needed

documents the prompt forbade -- flagging the conflict rather than substituting something adjacent. The review protocol is reproduced at Appendix I so that the design can be criticized rather than taken on trust.

The review is not independent of substrate. It is independent of position. §8 argues that this distinction is the paper's most practically consequential finding; here it is a limitation, and it is stated as one.

3.5. The confound that is conceded and not dissolved

The four production conditions described in §4 are sequential. Learning is confounded with the governance variable throughout, and a reader may explain the entire trajectory as the human operator becoming more skilled at instructing the system. The evidence cannot exclude that explanation, and the paper does not claim to.

What the confound does not touch is §5. Every finding there is a single-condition intervention on a fixed artefact: the same instrument, the same seed, one element removed. Ablation results are not trajectory claims and do not inherit the trajectory's confound.

3.6. Reproducibility

The instrument is deterministic. The source is content-addressed and every citation is line-numbered. The ablations and the reviewer's probes are wrapper scripts over unmodified exported functions. A third party given the artefact set reproduces exactly those figures this paper marks RETAINED, and Appendix III lists them with their digests.

The distinction is load-bearing and is stated here rather than in a footnote. A figure is RETAINED when the artefact behind it is released and the check re-runs; it is UNRETAINED when the instrument that produced it no longer exists, so that what survives is this system's attestation about its own past output. Under the rule at §10.3, an unretained computed figure may be reported but may not be counted as support.

Applying that rule to this paper withdrew three of its quantified consequences -- the conditions table at §4, and the ablation figures at E1 and E3 -- and Appendix III records what was attempted, what re-ran, and what did not. A reader who wants the shortest route to the paper's weakest point should begin there. The artefacts are held under a confidentiality constraint described in §4; the constraint bears on what can be published, not on what can be verified by a party granted access.

4. The setting

The evidence comes from the production of a software architecture. What follows is the shape of that work, at the level of abstraction the argument requires. Its content is under a confidentiality constraint, and every example below is a governance measurement rather than an architectural feature. The abstraction is principled rather than evasive: no claim in this paper depends on what the architecture does.

The object. A formal core of about one hundred and sixty lines states, in ordinary prose, the small set of entities an architecture admits and the relations among them. Beneath it sit three documents in sequence: a design document derived from the core, an architecture derived from the design, and a specification in six parts derived from the architecture. Each is derived from the one above. A defect introduced at any level propagates downward and is not repaired by correcting the level beneath it.

The participants. One human operator, sole author of the formal core and of every governing decision; one artificial system serving as builder of record, which writes code, produces documents and runs verification. The division follows the three-register model: the operator performs the operations requiring perception of configurations as wholes, the system performs the sequential and decompositional operations, and a governed exchange moves material between them (Heller, 2026a, §4; Heller, 2026b, §7).

The instrument. The system built an executable apparatus encoding the formal core as rules and testing documents and code against them: a suite reporting its own validity, a build gate failing on a missing declaration, mutation tests requiring a deliberately broken case to fail by name. The instrument matters to this paper because it is the mechanism through which readings became data, and because it is the control that failed.

Four conditions. All four attempted the same stage of the same task with the same participants and object. What varied was what the acceptance criteria were bound to.

	Criteria bound to	Outcome
A	Nothing external	The system identified genuine gaps and closed each by authoring a constraint of its own devising. The specification roughly doubled in size, and the larger part of the result was system-authored text. The verification gate stayed green, because it proved conformance to the core's type rules and cannot see authorship. The specification was reset

	Criteria bound to	Outcome
B	Criteria the system authored	Its own criteria -- auditability, provenance, judgment removal -- were satisfied in full and mechanically enforced. Measured against a file of known-defective statements maintained by the operator, the instrument reached a small minority of them, and reported itself valid at the same moment. Separately, the system replaced semantic reading of the core with pattern-matching and described the substitution as *judgment removal*
C	Line-numbered source, verbatim-verified quotation, mutation proof	The derivation from the core was run to closure and re-expressed as an executable check, so that it could be refuted rather than believed. No question was referred to the operator
D	The same, plus files the operator maintained	The prior design document was inventoried at sentence granularity, every quotation machine-verified verbatim before any tool ran, compared mechanically against the derivation, and re-derived. The findings of §5 are all from the artefacts of conditions C and D

Every figure this table once carried has been withdrawn. Earlier drafts gave counts for conditions A, B and D: constraints authored, bytes before and after, defective statements reached, provisions inventoried, quotations verified. The runs behind them are not retained, and under the retention rule at §10.3 a computed figure whose instrument is gone may be reported but not relied on. They were not even reported here, because none of them is load-bearing -- what conditions A, B and D establish is qualitative, and the qualitative account above is the operator's contemporaneous attestation, marked as such.

No claim in §5 through §11 depends on a withdrawn figure. That was established by reading each of those sections against the withdrawn list rather than by recalling that it was so, and the check is repeatable: the figures appear nowhere else, which Appendix III records. What the table loses is rhetorical weight, which is the correct thing to lose -- a number that cannot be checked persuades exactly as well as one that can, and that is the property this paper exists to describe.

One clarification that prevents the obvious misreading. In condition B the system's criteria were not vague. They were written down, argued for, and enforced

by code that failed when they were violated. They were rigorous and they were met. What they lacked was a term whose value the system did not determine. That is the variable this paper isolates, and it is not the same variable as care.

5. The evidence

Five findings. Each is given in the same four parts -- the claim as the artefact states it · what it cites · what the cited source says · what happened under ablation -- because uniform structure is what separates evidence from anecdote.

E1 · An encoded relation whose target its cited line does not contain

The claim. A layer of the architecture is *forced* at a determinate step of the derivation, by the clash of two modes of a framing text.

What it cites. One line of the formal core, quoted verbatim, plus two lines defining the layer's signature and scope.

What the line says. It says that the clash "differentiates an ecosystem." The word naming the layer appears at four other lines of the core; the word *ecosystem* appears at four lines; no line of the core connects them, and one of the four ties ecosystems to a different entity entirely.

Under ablation. UNRETAINED. The figure this paper carried -- that removing the relation leaves the derivation closing over nineteen of twenty-two elements, reaching neither the layer nor the two elements beneath it -- was computed on a walk that is not in the released tree, and cannot be re-run. It is withdrawn (Appendix III).

What remains is retained and sufficient for the finding. The encoded edge names the layer as its target; the cited line does not name the layer; no line of the core connects the clash to it. All three are readable in the released artefacts, and the third was established by a party with no stake in the answer. Removing the relation therefore leaves the instrument with no encoded warrant for that layer -- which is the claim E1 needs. The count it does not need, and no longer has.

What it makes the flagship case. The encoding carried a claim field stating a proposition the line *does* support -- that a second entity has two modes. The edge's target was a different proposition, and only the claim was ever checked. The mechanism that verified the encoding checks that the quoted text is present at the cited line. It does not check that the line carries the proposition the edge asserts. That limitation was written down, in the file, by the system that then relied on the mechanism as though it were not there. No artefact discloses this encoding as a reading.

E2 · A relation the code correctly refused to compute with, which the prose stated as established

The claim. A particular entity is the holder of the first element the derivation forces.

What it cites. Two lines of the core, and a check of the instrument.

What the sources say. The first line says the element is held by *a member*. The cited check tests only that a member-typed element is the first forced; it says nothing about which entity that member is. The identification is carried by a separate relation in the instrument, marked as an inference, and the instrument's own rule states that inferred relations "are not edges... reported as findings, never used to compute."

Under ablation. Not applicable: the relation was correctly excluded from the computation.

Why it is the sharpest structural instance in the set. The mechanism worked exactly as designed. The provenance discipline held in the code and was lost in the transfer to prose. A classification that governs data and does not travel with the claim into the document governs nothing that a reader will see.

E3 · A policy stated in the code as its own opposite

The claim. Three elements of the architecture arrive in a strict order, and two structural claims of the design document rest on that order.

What it cites. The instrument's derivation.

What the instrument does. The code's own comment states the principle: where one element in the frontier is required by another, "the primitive has ordered them and the choice is not open." The code applies that principle only in the single case where exactly one minimal element exists; in every other case it adds all elements at once, dependents included.

Under ablation. UNRETAINED, and the re-run reported either way. The figure this paper carried -- that applying the principle consistently yields a materially different ordering, costing two structural claims and a pass condition -- was computed on a walk that is not in the released tree, and is withdrawn (Appendix III). Re-run on the walk that *is* retained, from the same seed, with the stated principle applied at every step, the ordering does not change at all.

E3 stands on its retained half, which is the half that matters. The discrepancy is between a comment and the code beneath it: the comment states the principle without condition, and the code applies it only where exactly one minimal element

survives. Both are readable in the released instrument, and neither requires a run. What is withdrawn is the claim about what the discrepancy costs -- which is precisely the kind of consequence-claim this paper argues should never outlive its computation.

What it shows. The ordering is partly an artefact of an unstated policy of the walking algorithm, and the code states that policy as its negation.

E4 · A claim cited to a source containing neither of its terms

The claim. A stated arity is a *minimum*, on the ground that a relation among two parties is reciprocity while a relation among three is obligation.

What it cites. A line of the core.

What the line says. It is a row of a signature table containing the arity and nothing else. The words *minimum* and *reciprocity* appear nowhere in the core. The instrument's own report marks this arity undetermined and prints that verdict in the same run.

Under ablation. Not applicable; the claim is prose. It is included because the system's report and the system's document disagree, in the same session, without either noticing.

E5 · Explicit provenance circularity

The claim. A placement decision, recorded as resolved.

What it cites. A governing document, amended the same day to state that placement.

What the amendment cites. The resolution.

Under ablation. Not applicable. The loop is visible in the citations and was not seen by either party at the time.

E6 · The disconfirming evidence, at full weight

Reported here rather than in a limitations section, because the finding it produces is the one that makes the remedy precise.

The tables transcribed from the formal core are the core. They match it line for line, and they survive every mutation applied to them. Forty-two of forty-two quotations verify verbatim at their cited lines. The bulk of the design document's added text is what it claims to be: the core's own words, the instrument's step numbers, and citations that resolve.

Every defect in the set sits in one layer, and nowhere else: the layer at which a reading became an edge. Transcription is clean. Execution is clean. Interpretation, encoded and then executed, is where all of it lives.

5.1. A cheap screen that appeared to corroborate the review, and did not

After the review, a mechanical screen was written over the encoded relation set. For each relation it asks two questions: is the quotation verbatim at its cited line, and is any surface form of the edge's target present at that line at all? The second is not a test of whether the line carries the claim -- that is a reading -- but of a necessary condition for it to do so.

The screen returns: forty-two of forty-two quotations verbatim; thirteen of fifteen edges with both ends lexically present; two flagged. The two flagged are E1 and the disclosed reading underlying part of E3's ordering.

That was reported, in an earlier draft of this paper, as independent corroboration of the review at trivial cost. It is not, and the correction is the most useful thing in this section.

The first half stands. The verbatim check is mechanical: it normalizes whitespace and emphasis and asks whether the quoted string occurs at the cited line. Forty-two of forty-two is a fact about the artefacts and re-runs today.

The second half does not. Deciding whether a target is "lexically present" requires knowing which words count as surface forms of that target, and the screen answers from a hand-authored table of twenty-five entries carrying thirty-one surface forms -- written by the same party that wrote the relations it screens, and disclosed in no draft. Ablating that table dissolves both findings:

Surface Table	Target-Absent	Source-Absent
As written	The relation at E1	The relation at E3
Target's entry widened by one word present at the cited line	None	The relation at E3
Source's entry widened by one word present at its cited line	The relation at E1	None
Both widened	None	None

The words are not exotic. Each flagged relation is flagged because the table declined a word that its cited line actually contains.

What this means for the claim that the screen was given no information about the review. True of its inputs, false of its construction. The table was authored by a party who knew the findings, and the information entered there. A reading was encoded as data, the data produced the expected output, and the output was reported as independent corroboration of the reading. That is the cycle at §2.1, in four sentences, committed by this paper in the section where it was most confident -- and it survived the drafting, the review's exclusion list, and every check described at §6.

Two things follow, and the second is the useful one.

First, §5's E1 and E3 do not depend on this screen. Their ground is the independent review's reading of the cited lines, and, for E1, the inspectable fact that the encoded edge names a layer while its own claim field states a different proposition. The screen was corroboration, and the corroboration is withdrawn.

Second, the defect was found by ablating the screen's judgement surface, and the run took one command. The wording matters: a surface table is not an encoding in this paper's sense -- it is the instrument's own vocabulary, not a reading rendered as data -- so ablating it is the same *method* applied outside the population the remedy at §10 defines. The remedy as stated would not have reached this. That gap is why the companion protocol carries a separate control for judgement surfaces. A control the paper recommends caught the paper. That is the strongest evidence available that the control works, and it was obtained at the paper's own expense, which is the only way this particular evidence can be obtained at all.

The full data, the surface table and this ablation are at Appendix IV.

6. The verification that could not see any of it

Throughout conditions C and D the instrument reported itself valid. Three structural properties of that report generalize beyond this instrument.

The validity flag excluded the checks that were failing. The aggregate flag was a conjunction over a named subset of checks. Two checks outside that subset were failing -- one of them refusing a piece of built code -- and both were printed as findings beneath the flag that declared the system valid. No inspection of the flag reveals this; it requires reading the line of code that composes it.

The provenance counter was a whitelist. The instrument reported "hand-declared: zero", meaning that no encoded relation fell outside a set of provenance labels. The

labels were chosen by the system that wrote the relations. The count is true by construction and carries no information about whether the readings are sound.

The documentation gate checked existence, not currency. A build gate required every unit to ship documentation and failed when documentation was missing. The instrument's own documentation stated that a capability was untested which had by then been built, and stated a count of checks that was no longer correct. Both passed the gate, because the gate asks whether a document exists.

The generalization. Wherever a system reports its own validity, three questions are prior to the report: *what does the flag conjoin, and what does it exclude · who chose the labels any count ranges over · is the description of the system current with the system*. None requires access to internals. All three were missed here across four sessions by parties actively looking for exactly this class of failure.

7. What the borrowed warrant is not

The mechanism resembles several better-known phenomena. It is none of them, and each distinction carries a test rather than an assertion.

Neighbour	Why it resembles	The distinguishing test
Goodhart's Law	A measure ceases to inform	Goodhart requires optimization pressure on the measure. Here the measure was satisfied correctly, under no pressure, and still carried No information. *Test:* remove every incentive to score well. The mechanism persists
Circular reasoning	The loop	No proposition is assumed in its own proof. The loop runs across artefact *types* -- reading, datum, execution, report -- and is invisible within any one type. *Test:* examine each link in isolation. All are valid
Confirmation bias	A prior survives	Requires a preference for the conclusion. The encodings were ones the author had no stake in and would have altered on request. *Test:* ask whether the conclusion was preferred. Here it was merely earlier
Automation bias	Trust in machine output	That is a human over-trusting a system. This is a system over-trusting its own prior

Neighbour	Why it resembles	The distinguishing test
		artefact, and the human's trust is downstream of a report that was accurate
Model collapse	Own output re-consumed	That is statistical degradation across training generations. This is provenance laundering within a single task, with no retraining, and with fluency undiminished
Ordinary error	A mistake	An error is corrected when found. This survived every check designed to find it -- built for that purpose, by the party that committed it

The classification that follows. The borrowed warrant is a provenance failure, not an epistemic or a motivational one. Remedies aimed at reasoning quality, incentive design, or diligence do not touch it, and §10 is constrained accordingly.

8. The twin relation

	Borrowed voice	Borrowed warrant
What is borrowed	Vocabulary	Warrant
Borrowed from	The request	The system's own prior artefact
What is masked	An operation	A reading, wearing the form of a result
Why invisible	The vocabulary came from the request	Every surface criterion passes
Detection requires	A different register	A different position
Cost of detection	A second kind of intelligence	A second instance of the same kind
What accumulates	A degraded output	A substrate later work derives from

Two consequences follow, and they point in opposite directions.

The borrowed warrant is the cheaper failure to control, and it is not controlled. The remedy for the borrowed voice is an external map of request patterns known to

exceed the system's register, built up from detections that only a different kind of intelligence can make (Heller, 2026a, §6). That is expensive and irreducibly human. The remedy here is a second instance of the same architecture, a fixed question set, and an ablation harness -- available to any practitioner today, at the cost of one additional run.

It is the more consequential failure to leave uncontrolled. A borrowed voice yields a bad output, and the damage is bounded by that output's use. A borrowed warrant yields a bad foundation -- and every artefact built on it is individually correct, individually well-cited, and individually defensible. The damage is bounded by the depth of the derivation tree, not by one document's reach.

9. Masking at scale

Paper I described a compounding risk running through *the reader*: as systems become more fluent, output that crosses the capability boundary becomes harder to detect, reliance grows, and the human capacity to detect boundary crossings degrades through disuse (Heller, 2026a, §8).

The borrowed warrant compounds through *the record* instead, and the two paths are independent.

Each cycle raises a reading's standing. Later work cites the raised standing rather than the reading. After three or four artefacts the encoding is unreachable behind a chain of correct citations, and recovering it requires not scepticism but re-execution with the encoding removed -- an operation nobody performs on a document that cites a passing test suite. The record hardens; the reader need not soften at all.

Where the risk concentrates. Any programme in which a system builds the instrument that verifies its own output. That is not an edge case: it is the default in agentic software production, in machine-assisted formal methods, and in automated evidence synthesis, wherever the same system is asked to do the work and to build the harness that checks it.

The corollary, stated rather than softened. Under this mechanism, additional verification infrastructure built by the same system is not neutral. It is a multiplier -- each new instrument is another site at which a reading can be encoded, executed, and cited back. The intuition that more checking is safer does not hold here, and the evidence of §6 is that the checking apparatus was the least examined part of the whole system.

10. Remedy

10.1. The negative result

Binding acceptance criteria to an external referent is necessary and not sufficient, and the evidence is unusually clean on this point because the failing derivation was bound throughout: every quotation line-numbered, verbatim, machine-verified, forty-two of forty-two.

The binding was to the citation's existence rather than to the proposition drawn from the cited text. Every defect in §5 lives in that gap. A mechanism that verifies the first while its documentation implies the second is **worse than no mechanism**, because it converts an open question into a settled one.

10.2. The operative form

The referent must be bound to the proposition, not to the reference.

Mechanically: a claim drawn from a source is supported only when ablating the source causes the claim to fail. §5.7 reports an attempt at a cheap approximation of this -- a necessary-condition screen over targets -- and reports its failure: the screen flagged the two relations a full review had identified, but only relative to a hand-authored surface it did not disclose, and its flags do not survive that surface's ablation. The lesson is not that cheap screens do not work. It is that a screen with a judgement surface is an encoding like any other, and inherits the same obligation: declare it, and ablate it. A screen that does both is still cheap.

10.3. Five controls

Stated here at principle level; operationalized as gates with negative controls, coverage reconciliation and stop conditions in the companion protocol. That reference is a courtesy and is marked as one. The companion protocol derives its controls from this paper, so this paper may not ground anything in it -- the two would then cite each other, which is the loop the third control below prohibits. Both documents derive from the definitions stated here, and neither stands on the other.

Provenance classes that travel into prose. Five classes -- transcribed, read, encoded, computed, external -- one per claim, carried inline in documents and not only in data. E2 is the case for the words *into prose*: the discipline held perfectly in the code and was lost at the boundary of the artefact type.

And a second axis alongside the class: retention. A claim is RETAINED when the artefact behind it is released and the check re-runs, and UNRETAINED when the run is

gone and what remains is the system's attestation about its own past output. This is not a sixth class -- the five are exhaustive and stay so -- it is a second question asked of every claim, and the two axes are independent: a COMPUTED claim may be either.

The rule it carries: a COMPUTED/UNRETAINED claim may not be counted as supported. It may be reported, with its status; it may not be relied on. The distinction matters because COMPUTED carries an implicit promise of re-execution, and a figure whose instrument has been deleted honours no such promise while reading exactly like one that does. §4 and §12 of this paper record what applying the rule here cost it.

Ablation as the standard of support. Untested encodings are recorded as unsupported, never as established. A load-bearing encoding that no document discloses is a stop condition.

The loop prohibition. No artefact may cite as its ground a document that cites it. E5 is a two-step loop that neither party saw; longer loops are correspondingly harder.

Validity composition printed at the point of verdict. Any aggregate validity flag prints what it conjoins and what it excludes, adjacent to the verdict. A green flag printed above failing checks is prohibited.

Judgement surfaces declared and ablated. Any screen or gate whose verdict depends on a hand-authored surface -- a vocabulary, a table of surface forms, a whitelist of permitted labels -- declares that surface where the verdict is reported, and ablates it. A verdict that moves when the surface is widened by one plausible word is a property of the surface, not of the source. This control is owed to §5.7's screen, which had such a surface, disclosed it nowhere, and does not survive its ablation -- while a second instrument in the same programme declared its surface in the file and ablated it as a gate condition. The discipline existed; it was not carried across. That is the ordinary shape of the failure: not ignorance of the control, but its uneven application.

10.4. The fourth governance mechanism

The three mechanisms these Papers have established all govern the boundary between registers: the alerting mechanism flags requests that cross it (Heller, 2026a, §6); the Three-Test Framework governs what may cross it as a contribution; the translation loop moves material across it in both directions (Heller, 2026b, §§7--8).

Positional independence governs something prior: the relation between a system and the standard by which its work is judged. Where that standard is unbound, or bound to an artefact of the system's own making, the other three mechanisms operate on a criterion the system has already made satisfiable. It is not a register, not

a second opinion, and not peer review. It is the same capability, applied by a party with no stake in the answer -- and on this evidence it reached, in one pass, defects that four sessions of self-applied checking had passed over.

11. Prediction and falsification

The paper's first prospective confirmation, stated before the predictions because it bears on the central asymmetry. The test asked whether the detection reported at §5 was owed to the reviewer's independence or to the pre-registered question set it was given. It has now been run.

The design. One variable was manipulated and the rest held: the same question set verbatim, the same primary artefacts, the same architecture, the same instruction that finding nothing was an acceptable outcome -- and the authoring party's own account, excluded by name in the original review, supplied in full. Four benchmark defects from the original return were fixed in advance, with the threshold and the consequence of each outcome settled in writing before the return existed.

The result: none of the four was found. Two were reported affirmatively absent after the reviewer read seventy-five cited source lines one by one; the other two are named nowhere in the return. Under the rule fixed in advance, this is the registered negative outcome: the exclusion of the authoring party's account was load-bearing, and the question set does not account for the detection.

What it establishes and what it does not. It is one run against one artefact set, so it fixes no rate. It isolates access to the authoring party's account, which is narrower than the word *position* covers, and it is the narrower thing the remedy at §10.4 actually turns on. What it removes is the alternative explanation, and that is what a falsification is for.

The pre-registration and the return are both retained, because a computed result whose instrument is discarded may not be counted as support (§10.3).

Given a task whose success is measurable against an external standard, a system permitted to encode its own readings of a source and to cite executions of those encodings as evidence for the readings will produce load-bearing conclusions whose survival under ablation is uncorrelated with the source's support for them.

How to falsify it. Any instrumented derivation with an ablation harness suffices. Class each encoded element as transcription or reading; ablate each; record what fails. If readings ablate no differently from transcriptions, the mechanism is not present. Appendix II gives the harness design; §5.7's screen gives a cheaper first pass.

A second, weaker prediction, offered because it is nearly free to test. Defect concentration will be found at the encoding layer -- where a reading becomes a datum -- and not in transcription or in execution. In this case the concentration was total: thirteen of fifteen edges mechanically clean, forty-two of forty-two quotations verbatim, and every finding in the remaining two edges and the prose derived from them.

12. Limitations

Stated in the order a reader will raise them.

One dyad, one model family, four sequential conditions, with learning confounded with the governance variable throughout (§3.5). The ablation results do not inherit this; the trajectory claims do.

The analytical instrument was built by the system whose failures it reports, and the first interpretation of the evidence was that system's. The adversarial review addresses the paper's central question and not the rest.

The review was conducted by the same architecture. It achieved independence of position, not of substrate -- which is the paper's claim, and therefore also its bound.

Position was confounded with question specification. The confound was tested and did not survive. §3.4 lists three differences between the reviewing and the authoring party: no stake in the answer, a clean context, and a question set fixed before the reviewer saw any artefact. The paper attributes the detection to the first; it could as well have been owed to the third, and the original design did not separate them.

A second review was therefore run with the question set held identical and the authoring party's account supplied instead of excluded. It found none of the four benchmark defects. §11 gives the design and the result. Specification does not account for the detection, and this limitation is discharged rather than merely declared.

What replaces it as a limitation is narrower and remains open. The test isolates access to an account, not stake: the reviewing session did not write the artefacts and had no investment in defending them. Whether having produced the work degrades detection further is untested, and if it does, this paper understates the effect rather than overstating it.

One observation from the same run is recorded here as a hypothesis and is claimed nowhere in this paper. The two reviews returned largely disjoint defect sets, and the second's most consequential finding was reachable only because it held the record the

first was denied. That suggests account-access may select between classes of defect rather than simply degrade detection -- but the benchmark set contains no defect of the second class, so this design cannot measure it. Settling it needs a second pre-registered benchmark set and two further runs, and until those exist the observation supports nothing.

The measures are counts, not scores, and are not commensurable with the five-point scale used in Paper III.

The failure record is maintained by the system that failed, so failures neither party recognized as failures are absent by construction. This is the most serious limitation in the list, because it is the one the mechanism itself predicts.

Disclosure constraints mean the object is described in shape only. A reader cannot independently assess whether the abstraction conceals a relevant particular, and must take the author's word that no claim depends on the omitted content.

Three of this paper's quantified consequences were computed on instruments that no longer exist, and were carried forward in prose after the instruments were gone: the conditions table's figures at §4, and the ablation figures at E1 and E3. All three are withdrawn, and Appendix III records what could and could not be re-run. The paper set out to publish a complete audit record of its own evidence and could complete it only where an artefact outside the session survived. This is §9's mechanism -- the record hardening while the computation behind it disappears -- operating on this paper, and it was not caught by any of the checks built to catch it.

The screen reported at §5.7 was found, after the review closed, to rest on a hand-authored surface table that no draft disclosed, and not to survive that table's ablation. It is the defect §6 attributes to the instrument, committed in the paper's own apparatus, in the section where the paper was most confident. It was found by checking a line count, not by insight -- which is the least flattering way to have found it and the most instructive.

Appendices

I -- The adversarial review protocol, including the exclusion list and the instruction that finding nothing is an acceptable outcome. Every operative instruction is reproduced without alteration; the source list and document names are the architecture's own and are replaced by descriptions of their class, with each redaction marked where it falls. The word *verbatim* stood here in an earlier draft and is withdrawn: a redacted text described as verbatim is this paper's own subject in miniature.

II -- The ablation harness design: classing, removal, re-run, and the load-bearing / contributory / inert classification.

III -- The audit boundary: what re-runs from the released artefacts, with digests and commands; what does not, with the reason; and the outcome of every re-run attempted for this revision, including the two that failed. The appendix promised here in an earlier draft was the four-condition record. It could not be written, and why it could not is the subject of the appendix that replaced it.

IV -- The encoded relation set, every entry classed, with the mechanical screen's output per entry: quotation verbatim at line; edge source lexically present; edge target lexically present. This is the paper's primary data and the first thing a sceptical reader should attack. Element identifiers are pseudonymized: the relation set's names are the architecture's own taxonomy and are withheld under the disclosure constraint of §4. The substitution is deterministic and structure-preserving -- every relation, every flag, every count and every surface-form total survives unchanged, and only the names are opaque. The mapping travels with the artefact set and not with the paper (§3.6).

Appendix I -- The adversarial review protocol

Reproduced as issued. The source list and the document names are the architecture's own and are replaced by descriptions of their class; every operative instruction is reproduced without alteration. The redaction is marked at each point where it falls.

What you are. You are the reviewer of record for a bounded provenance review. You are not the builder. You will not be asked to fix anything, build anything, or write prose for any document. Your entire output is a set of findings in the form this prompt specifies, and it will be adversarially triaged by the builder of record afterwards. No text you write enters any project document.

What you must not read, and why. Do not read, search for, or rely on any of: *[the builder's session summaries for the two sessions under review; the running session log; the hand-off note; the resume capsule; any document whose title contains* evaluation, capability boundary *or* retest*]*.

These carry the builder's own account of what it did and why. You are reviewing that account's substance against the primary record. A reviewer who reads the account first will confirm it fluently, in its vocabulary, and the review will be worthless. If you find yourself about to open one of these to "understand the context", that is the failure this instruction exists to prevent: the context you need is in the artefacts themselves.

If you cannot answer a question from the primary sources alone, say so and stop on that question. An unanswerable question is a finding. A guess is not.

The primary sources -- read these, in full, at depth: *[the formal core, 160 lines, from which everything is supposed to derive; the derivation from the core, step by step; each derived element tied to an existing specification or recorded as a demand; the prior design document inventoried and compared against the derivation; the re-derived design document as installed; the source snapshot carrying the instrument; the settled record the grounding cites]*.

Restore the code to disk from the snapshot and run it. The snapshot carries its own extractor. Claims about what an instrument does are checkable by running it, and a claim you have only read is a claim you have not checked.

The three questions. Answer each independently. Do not let your answer to one shape your answer to another.

Q1 -- Provenance of the ordering. The re-derived design document is organized by a particular ordering of elements. Establish where that ordering came from. State the earliest artefact in which that ordering appears, the operation by which it got there, and whether that operation reads from the formal core or originates with whoever performed it. Quote the ordering as it appears in the earliest artefact and as it appears in the installed design, and say whether they are the same ordering. If the ordering was produced by executable code, run that code and report what it emits. If the emitted order differs from the document's, that difference is the finding.

Q2 -- Unsourced content (the load-bearing question). The claim under review is: no mechanism, invariant, constraint or rule entered the specification without tracing to a cited source. Test it adversarially. Go through the installed design document and the three artefacts looking for any mechanism, invariant, constraint, rule, threshold, arity, ordering or prohibition that is stated as though established, and does not trace to the formal core, the settled record, or a named prior specification, and is not explicitly marked as a demand, an inference, or an open finding. Report every instance with its location and the search you performed to try to source it. Report the count. Zero is a permissible answer only if you actually searched -- state how.

Note the trap: a citation that exists is not a citation that supports. Where a claim cites a line, read that line and say whether it carries the claim. A cited line that does not carry its claim is a stronger finding than an uncited claim.

Q3 -- What the instrument does not reach. The instrument reports itself valid. Establish what that validity does and does not cover. Identify claims made in the artefacts or the installed design that no check in the suite tests, and say for each whether it is testable-but-untested or untestable-in-principle. Run the suite; report its own statement of validity; then state plainly what a green suite here is evidence *of*.

How to answer. One artefact, structured as: **Q1** -- the ordering's provenance, with the earliest artefact, the operation, and your run output · **Q2** -- a table: location · the claim · what it cites · whether the citation carries it · your verdict · **Q3** -- the untested set, split testable / untestable, with the suite's own output · UNDETERMINED -- every question or sub-question you could not settle from the primary sources, with what you would have needed · YOUR OWN CONFIDENCE -- per question, and what would change it.

State disconfirming evidence with the same prominence as confirming evidence. If the answers are dull -- the ordering is sourced, nothing is unsourced, the coverage is as stated -- say so plainly. A review that finds nothing and says so is worth more than a

review that manufactures a finding, and this one is being read by someone who expects it might find nothing.

Do not recommend, do not repair, do not propose next steps. Findings only.

Three properties of this prompt did the work, and they are the transferable part. The exclusion list removes the builder's own account from the reviewer's reach, because a reviewer who reads that account first confirms it in its own vocabulary. The run instruction converts claims about instruments into claims checkable by execution. The permission to find nothing, stated twice, removes the incentive to manufacture a finding -- and it is the sentence that makes a null result informative rather than embarrassing.

Appendix II -- The ablation harness design

The principle. A claim resting on an encoded reading is supported only if removing the encoding causes the claim to fail. The harness therefore does one thing: it removes one encoding at a time and re-runs the computation that the claim cites.

The construction. Every ablation is a wrapper over the instrument's unmodified exported functions. No instrument file is edited. The relation set or the surface table is filtered in memory before the graph is built, and the same entry point is called with the same seed. This matters for a reason the paper's own subject makes obvious: an ablation performed by editing the instrument cannot be distinguished, afterwards, from an ablation that changed something else as well.

The classification. Each ablated encoding is recorded as load-bearing (removal changes the result), contributory (removal changes an intermediate but not the result) or inert (removal changes nothing). An encoding that is inert has no evidential standing whatever, and an encoding that is load-bearing and undisclosed is a stop condition.

A limitation of this harness, stated because it runs in the direction that flatters the result. The procedure above removes one encoding at a time. Where two encodings redundantly carry the same reading, removing either alone changes nothing and both class as inert -- and the reading is then recorded as unsupported-and-harmless when it is in fact doubly supported and never examined. The harness under-reports by construction, and every figure it produces is a lower bound.

The requirement that closes it: encodings sharing a target are ablated jointly as well as singly, and an encoding classed inert singly but load-bearing jointly is recorded as load-bearing. This was not done for the ablations reported in this paper, which

were single-element throughout; the results stand as lower bounds and are marked here rather than in a footnote.

The three ablations this paper reports, and their standing.

Ablation	Instrument	Outcome
Surface-form ablation of the derivation's declared judgement surface	Retained; runs today	2 of 21 forms load-bearing; origin scoping load-bearing
Surface-form ablation of the screen at §5.7	Retained; runs today	Both of the screen's findings are surface-dependent and vanish under single-word additions
Removal of the single encoded relation at E1	Not retained	The figure the paper cited cannot be re-run; see Appendix III
Consistent application of the tie-break principle at E3	Retained; runs today	No difference on the retained walk; the original figure was computed elsewhere; see Appendix III

What this table is for. Two of the four ablations the paper relies on do not stand as originally reported, and both were found by attempting to rebuild this appendix. A harness described but never rebuilt is an assertion. The rebuilding is the control, not the description.

Appendix III -- The audit boundary

What re-runs today, from the released artefacts. Each is content-addressed; the digest is the leading sixteen hexadecimal characters of the SHA-256 of the file as released.

Artefact	Bytes	Lines	Digest
The formal core	13,219	160	a11e36a5cf7566ad
The relation set	15,586	198	2ca72ff3d6bd8292
The derivation	22,614	374	b5f1d792fee94ec0

Artefact	Bytes	Lines	Digest
The cross-frame check	19,882	334	cbe43baa120767f3
The suite	18,303	243	7fa66cf987828c0e
The engine	27,304	463	185d352faa50c96c
The screen of §5.7	3,592	56	48182130de593925

Every figure in §5 that survives is recomputed from these, by running the suite and the screen. The suite reports its own validity, the frontier's derivability and the cross-frame result; the screen reports forty-two of forty-two quotations verbatim and two flagged relations. Both were re-run at the moment this appendix was written.

What does not re-run, and why. The items below are transcribed from the independent review's own record of what it could not settle, together with two found while rebuilding Appendix II. They are listed, not reconstructed.

Item	What is missing	Consequence
The sentence-level inventory of the prior design and its verbatim checks	The working directory that produced it	§4's condition-D figures are withdrawn
The earlier scratch walks and the demand-register run	The same directory	§4's condition-C detail rests on the retained suite only
The frozen base, the marked redline and the byte-identical recovery	The architect's records, not the released set	Not cited in this paper
The conformance re-run over the design of record	The design is not in the released set	Not cited in this paper
The four-condition byte counts and constraint counts	No retained run	§4's conditions A and B figures are withdrawn
The articulation walk over twenty-two elements	The instrument is not in the released tree	E1's and E3's consequence figures are withdrawn

Two re-runs were attempted for this revision and are reported either way. The ablation at E1 cannot be run: the walk it cites does not exist in the released tree. The ablation at E3 runs and returns no difference on the retained origin walk, from the same seed, with the stated principle applied at every step. Both outcomes are

recorded here rather than in a footnote, because a failed re-run is the result the paper's own standard treats as decisive.

Why the absences exist. The missing items are session-scratch working directories, which the governing document-handling policy deliberately does not carry into the durable store: audit-class material stays with the party that produced it and does not accumulate in the shared record. The absence is a policy consequence, not an oversight -- and the policy is defensible for reasons that have nothing to do with this paper, which is why this entry is a disclosure and not a recommendation.

What the absence demonstrates. The paper set out to publish a complete audit record of its own evidence and could complete it only where an artefact outside the session survived. Every figure that could not be re-run had been carried forward in prose long after its instrument was gone, and none of the checks built to catch exactly this noticed. That is §9's mechanism -- the record hardening while the computation behind it disappears -- operating on this paper. It is reported here because the alternative was to publish the figures and say nothing.

Appendix IV -- The encoded relation set, every entry classed

Pseudonymized per the note above. ly denotes a layer of the architecture, kd a kind, md a mode, rg a region, xx an element outside those classes. Relation identifiers are unchanged. A law or note has no edge and therefore no source or target column.

ID	Line	Mark	Relation	from	to	Quote Verbatim	Source Present	Target Present
--	--	--	--	--	--	--	--	--
K1	13	STATED	law	--	--	yes	--	--
K2	19	STATED	requires	kd1	xx1	yes	yes	yes
K3	21	STATED	law	--	--	yes	--	--
K4	17	STATED	requires	kd2	kd3	yes	**NO**	yes
K5	18	STATED	requires	xx2	kd2	yes	yes	yes
K6	21	STATED	requires	rg2	kd3	yes	yes	yes
M1	30	STATED	requires	md2	kd1	yes	yes	yes
M2	29	STATED	requires	md1	rg1	yes	yes	yes
M3	32	STATED	law	--	--	yes	--	--
M4	34	STATED	enables	md1 +	ly1	yes	yes	yes
M5	35	STATED	enables	md1 +	ly4	yes	yes	
M6	36	STATED	enables	md3 +	ly5	yes	yes	**NO**
M7	38	STATED	law	--	--	yes	--	--
P1	44	STATED	mutual	md1	ly1	yes	yes	yes
P2	45	STATED	requires	ly1	md2	yes	yes	yes
P3	49	STATED	enables	md2	ly1	yes	yes	yes
C1	53	STATED	excludes	kd3	kd3	yes	yes	yes
C2	57	STATED	law	--	--	yes	--	--

ID	Line	Mark	Relation	from	to	Quote Verbatim	Source Present	Target Present
S1	61	STATED	law	--	--	yes	--	--
S2	71	STATED	law	--	--	yes	--	--
S3	73	STATED	law	--	--	yes	--	--
S4	73	STATED	law	--	--	yes	--	--
J1	77	STATED	law	--	--	yes	--	--
J2	82	STATED	law	--	--	yes	--	--
J3	83	STATED	requires	xx3	rg3	yes	yes	yes
J4	87	STATED	law	--	--	yes	--	--
D1	99	STATED	law	--	--	yes	--	--
D2	106	STATED	requires	xx4	ly2	yes	yes	yes
D3	108	STATED	law	--	--	yes	--	--
D4	108	STATED	law	--	--	yes	--	--
D5	110	STATED	law	--	--	yes	--	--
A1	113	STATED	law	--	--	yes	--	--
A2	127	STATED	law	--	--	yes	--	--
A3	129	STATED	law	--	--	yes	--	--
A4	133	STATED	law	--	--	yes	--	--
A5	134	STATED	law	--	--	yes	--	--
A6	141	STATED	law	--	--	yes	--	--
R1	150	STATED	law	--	--	yes	--	--
R2	156	STATED	law	--	--	yes	--	--
I1	30	INFERRED	note	--	--	yes	--	--
I2	36	INFERRED	note	--	--	yes	--	--
I3	82	INFERRED	note	--	--	yes	--	--

Totals

- Relations 42
- STATED 39
- INFERRED 3
- Laws 24
- Notes 3
- Edges 15

Surface Table

25 entries, hand-authored, forms per entry:

19 entries carry 1 surface form(s)

6 entries carry 2 surface form(s)

total surface forms: 31

How to read the two flagged rows. K4's source is not lexically present at its cited line; M6's target is not. Neither flag is a finding on its own -- lexical absence is a

necessary-condition failure, not a demonstration that the line lacks the proposition. Each was then read, and the readings are E1 and the disclosed reading underlying E3's ordering.

And the flags are surface-relative. Widening M6's target to admit one further word present at the cited line clears it; widening K4's source to admit one further word present at its line clears that. With both widened the screen reports nothing at all. The surface table is the control surface, and it was hand-authored by the party whose readings it screens. A reader attacking this appendix should attack the surface table first, and the paper's position is that they are right to.

References

Heller, N. A. (2026a). *The Capability Boundary Principle: Designing Intelligent Ecosystems Across Architectural Substrates.* Paper I of the Capability Boundary Papers.

Heller, N. A. (2026b). *The Capability Boundary Principle Applied: A Diagnostic Map.* Paper II of the Capability Boundary Papers.

Heller, N. A. (2026c). *Empirical Evidence from Governance Architecture Production.* Paper III of the Capability Boundary Papers.

Heller, N. A. (2026). *The Capability Boundary Protocol.* Operational companion to the Capability Boundary Papers; converts them into gates.

Paper IV of the Capability Boundary Papers · 17 August 2026